% Rotina para desenvolvimento dos modelos de PLS utilizando dados continuos

% e propriedades fisico-quimicas

% Essa rotna esta dividida nas seguintes partes:

% 01 - Conhecendo o plsmodel

% 02 - Aprimorando o modelo.

% 03 - Exemplo real.

% 04 - Praticando.

% Extras:

% 00 - Edicao de Figuras

% 00 - Edicao de Figuras da Monografia.

% 00 - Conhecendo a funcao for

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

%%

%%%%%%%%%%%%%% 01 - Conhecendo o plsmodel

% Recomendo comecar a pratica com os seguintes comandos.

clear % Limpa o Workspace

clc % Limpa o Command Window

close all % Fecha qualquer imagem aberta.

% Agora iremos puxar os pacotes necessarios. (OC Exclusivo do Octave)

pkg load statistics

pkg load io

% Vamos mudar o Diretorio para onde colocamos os dados do IFES-Ciencia

% {Lembre-se de substituir o diretorio dentro do comando "cd" pelo diretorio correto.}

cd('C:\Users\Pedro\OneDrive - aluno.ufes.br\Quimiometria\IFES Ciencia\PLS\_Model');

%cd('...\IFES Ciencia\PLS\_Model');

load('Dados\_API.mat')

% Note que os dados nos temos;

% Xmir % Espectro MIR das nossas amostras.

% y % Sao os dados quantitativos.

% objetos % E a separacao de calibracao e teste das amostras.

% Nmir % Comprimento de onda do espectro MIR.

% Assim, teremos que separar os dados entre calibracao e teste usando as

% seguintes linhas de comando.

Xcal = Xmir(objetos.cal,:); ycal = y(objetos.cal,:);

Xtest = Xmir(objetos.test,:); ytest = y(objetos.test,:);

% No plsmodel temos que utilizar um conjunto de opcoes para ele funcionar.

% E como se fosse um manual de instrucoes, ou mapa, para o pls fazer os

% calculos da forma desejada.

options=[];

options.Xpretreat = {'center'};

% Pretratamento da matriz Xcal;

% ycal sempre centralizado;

options.vene = 5;

% Tamanho da janela de validacao cruzada, recomanda-se 5;

options.vl = 20;

% Numero de variaveis latentes;

% Como nosso objetivo inicial e otimizacao colocamos 20 para ver como o

% RMSECV se comporta em cada variavel latente.

% O RMSECV e o erro quadradico medio de calibracao cruzada, este parametro de

% avaliacao e utilizado para podermos otimizar o numero de variavel latente.

% Vamos rodar o plsmodel utilizando so as amostras de calibracao para

% termos uma nocao do comportamento do RMSECV ao longo das variaveis

% latentes (VL).

modelo=plsmodel2(Xcal,ycal,options)

% Vemos no grafico gerado que a maior queda de RMSECV ocorre entre o VL 3 e

% 4, todavia o RMSECV aparenta estabilizar entre os valores 2,4 e 2,6,

% faixa primeiro alcancada pelo VL 4. Assim, utilizamos este valor.

options.vl = 4;

modelo = plsmodel2(Xcal,ycal,Xtest,ytest,options);

% No 'modelo' temos alguns parametros de avaliacao que merecem destaque:

modelo.RMSEC; % 2.3162

% Erro quadradico medio de calibricao, onde podemos ter uma nocao do erro

% das amostras de calibracao.

modelo.RMSEP; % 1.9387

% Erro quadradico medio de previsao, onde podemos ter uma nocao do erro das

% amostras de teste.

modelo.R2c; % 0.9414

% Coeficiente de determinacao das amostras de calibracao, quanto mais

% proximo a 1, melhor o modelo.

modelo.R2p; % 0.9440

% Coeficiente de determinacao das amostras de teste.

% Com estes quatro parametros podemos avaliar se o modelo e bom, ou nao,

% pelos altos valores de coeficiente de determinacao, poderiamos dizer que

% o modelo foi bem sucedido, entretanto, quando analisamos a mesma

% propriedades na literatura, percebemos que o RMSEP considerado bom e

% proximo ao 1.2 e o R2c acima de 0.95. O que podemos fazer aqui, para

% melhorar o modelo, e testar outros valores de VL e outros

% aperfeicoamentos.

% Agora, para fins de teste, vamos testar outros VL. Eu recomendaria testar

% o 5 e o 6, entretanto, tome cuidado ao escolher o VL, conforme tu

% aumenta o numero de variaveis latentes menor tende a ficar o RMSEC e

% maior tende a ficar o RMSEP, como explicado no artigo.

options.vl = 5;

modelo = plsmodel2(Xcal,ycal,Xtest,ytest,options);

modelo.RMSEC; % 2.0730

modelo.RMSEP; % 1.7584

modelo.R2c; % 0.9537

modelo.R2p; % 0.9500

options.vl = 6;

modelo = plsmodel2(Xcal,ycal,Xtest,ytest,options);

modelo.RMSEC; % 1.9061

modelo.RMSEP; % 1.4480

modelo.R2c; % 0.9614

modelo.R2p; % 0.9582

% Com base nos parametros de avalicao podemos dizer que o melhor modelo e o

% com seis variaveis latentes, devido a ambos os RMSE serem menores e ambos

% os R2 serem mais proximos de 1, entretanto, isso nao basta para afirmar

% que ele e melhor, devemos comprovar isso com alguns testes.

% Primeiro, fazemos os dois modelos que queremos comparar, um com 4

% variaveis latentes e outro com 6 variaveis.

options.vl = 4;

modelo4 = plsmodel2(Xcal,ycal,Xtest,ytest,options);

options.vl = 6;

modelo6 = plsmodel2(Xcal,ycal,Xtest,ytest,options);

%VL 4 6

%RMSEC 2.3162 1.9061

%RMSEP 1.9387 1.4480

%R2c 0.9414 0.9614

%R2p 0.9440 0.9582

% Assim, vamos comparar primeiro os modelos, avaliando por grafico de

% medido x predito, ou reference x predicted.

close all %Fechando imagens ja criadas.

subplot(2,1,1)

plot(modelo4.Ycal(:,1),modelo4.Ycal(:,2),'bo','LineWidth',1); hold on;

plot(modelo4.Ytest(:,1),modelo4.Ytest(:,2),'r\*','LineWidth',1); hold on;

ylim([5 65]); xlim([5 65]);

plot(xlim, ylim, '--k');legend('Calibration','Prediction','Location','southeast');

title('Modelo 4');

set(gca,'FontSize',12);xlabel('Reference','fontsize',12);

ylabel('Predicted','fontsize',12);

subplot(2,1,2)

plot(modelo6.Ycal(:,1),modelo6.Ycal(:,2),'bo','LineWidth',1); hold on;

plot(modelo6.Ytest(:,1),modelo6.Ytest(:,2),'r\*','LineWidth',1); hold on;

ylim([5 65]); xlim([5 65]);

plot(xlim, ylim, '--k');legend('Calibration','Prediction','Location','southeast');

title('Modelo 6');

set(gca,'FontSize',12);xlabel('Reference','fontsize',12);

ylabel('Predicted','fontsize',12);

% Caso queira melhor compreender estes comandos leia o 00 - Edicao de figuras.

% Para melhor avaliacao/comparacao, tente deixar os graficos na forma quadrada,

% afinal, tanto eixo x e y tem a mesma dimensao.

% Um grafico de medido e predito trata-se de uma comparacao sobre o real

% valor de propriedade dependente (y) de uma amostra e o que foi previsto

% pelo modelo. Entao, quanto melhor o modelo, mais proxima a amostra fica

% da linha de referencia.

% Na comparacao dos graficos feitos com VL 4 e 6, percebe-se que tem pouca

% diferenca na faixa 10 a 37 e uma diferenca significativa na faixa 40 a

% 60. Todavia isso, ainda, nao e o suficiente para determinar que o modelo

% 6 e realmente melhor.

% Teste de Acuracia.

% Depois, utilizamos o "accuracy\_test" que e uma funcao de comparacao de

% modelos multivariados, que utiliza testes randomicos, esta funcao ira

% confirma se existe uma diferenca estatistica.

%[pvalue,dist\_tt,meandiff] = accuracy\_test(yO,ypA,ypB,randuni,500000,0.05)

% yO ; Valor de refencia do y test.

% ypA ; Valor de y test previsto pelo modelo A.

% tpB ; Valor de y test previsto pelo modelo B.

% teste ; Tipo de teste, que pode ser:

% randbi : teste randomico bicaudal.

% randuni : teste randomico unicaudal. yhatA > yhatB

% niter ; Numero de permutacoes.

% alpha ; Nivel de significancia adotado;

%

% Neste caso iremos usar estes padroes.

tic

[pvalue,dist\_tt,meandiff] = accuracy\_test(modelo4.Ytest(:,1),modelo4.Ytest(:,2),modelo6.Ytest(:,2),'randbi',500000,0.05);

toc %{892 sec\_Pedro}

% Esta demorando? Nao se preocupe, costuma demorar mesmo.

% Tem duas formas de avaliar se os modelos sao diferentes;

% 1- Pela resposta da funcao, que caso sejam diferentes dara

% "Modelos com DIFERENï¿½AS na acurï¿½cia".

% 2- Comparar o valor do pvalue obtido com o alpha utilizado, caso o pvalue

% seja menor, os modelos tem diferenca estatistica e o modelo com 6 VL e

% melhor.

% Assim, temos duas confirmacoes que o modelo com 6 variaveis latentes ï¿½

% melhor que o modelo 4. Mas, isso e so o comeco e nem de longe o melhor

% modelos que podemos alcancar.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

%%

%%%%%%%%%%%%%% 02 - Aprimorando o modelo.

% Como visto na licao anterior, o PLS nao obteve um modelo bom em

% comparacao com o encontrado na literatura, entretanto, ainda nao

% extrairmos o maior potencial desta tecnica, porque podemos aplicar

% tecnicas de aperfeicoamento, como pretratamento, selecao de varaiveis e

% deteccao de outlier.

% Entao, vamos abordar alguma dessas tecnicas.

% Antes de comecar vamos deletar os dados da rotina anterior que nao

% usaremos.

clearvars -except ycal ytest Xcal Xtest objetos Nmir

% Este comando 'clearvars -except' deleta tudo exceto os arquivos citados

% em seguencia.

%% - Pretratamento,

% O Pretratamento, ou preprocessamento, tratase de uma modificavao da fonte

% analitica visando facilitar, ou aperfeicoar, a modelagem. Existem

% pretratamento que consideram somente a amostra isolada e que consideram a

% faixa da fonte analitica. Todavia, tem que tomar cuidado ao aplicar o

% pretratamento, pois ao mesmo tempo que tu remove ruido tu pode acabar

% removendo informacao.

% Para fazer o pretratamento, utilizaremos o funcao pretrata, que nos

% fornece os seguintes pretratamentos.

% {'center'} Centralizacao na media.

% {'auto'} Autoescalonamento dos dados

% {'snv'} Variacao padrao normal.

% {'msc'} Correcao multiplicativa de sinal

% {'deriv',[7,2,1]} Derivada

%

%[Xp,Xtp]=pretrat(X,Xt,{'auto'});

[Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'auto'});

% Voce deve escolher muito bem o pretratamento utilizado, pois caso utilize

% da forma incorreta podera esconder informacao e destacar desinformacao,

% prejudicando o modelo assim.

% Recomendo adicionar o numero "2" no espectro tratado para nao mistura a

% versao bruta e a versao tratada, alem disso, podemos ver se houve

% modificacao do espectro atraves do comando plot.

subplot(2,1,1)

% subplot(A,B,C)

% Esse comando permite colocar organizar figuras nas seguintes

% configuracoes, A linhas, B colunas, figura numero C.

plot(Nmir,Xcal);

subplot(2,1,2)

plot(Nmir,Xcal2);

% Perceba que o grafico mudou drasticamente de perfil, ficou praticamente

% irreconhecivel. O Autoescalonamento pode ser aplicado em dados de

% infravermelho, todavia, e mais recomendado para dados discretos de

% proporcoes distintas.

% Nos espectros de infravermelho e recomendado o uso de derivada.

%[Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[A,B,C]});

% O Tratamento com derivada tem certas configuracoes a serem otimizadas;

% A = Representa o numero da janelas, numero de variaveis que deve ser

% considerada para derivar, sempre recomendar numeros impares.

% B = Grau do polinomio, grau da derivada a ser utiliada, recomenda-se 1 e

% 2 para infravermelho.

% C = Ordem da Derivada, e quantas vezes a derivada sera feita, se e

% primeira derivada (1), ou segunda derivada (2).

% Eu, costumeiramente utilizo a seguinte configuracao.

% 15 Jenelas, 2 Grau de Polinomio, 1 Derivada

% Esta funcao, pretrat, foi desenvolvida para aplicar o mesmo pretratamento

% em dois espectros ao mesmo tempo, um de calibracao e outro teste,

% todavia, tu pode fazer uma 'jogada' para somente tratar um espectro.

[Xcal2,~]=pretrat(Xcal,Xcal,{'deriv',[15,2,1]});

% Mas, lembre-se, o Xtest deve ser OBRIGATORIAMENTE pre tratado em conjunto

% com o Xcal.

close all

subplot(2,1,1)

plot(Nmir,Xcal);

subplot(2,1,2)

plot(Nmir,Xcal2);

% Nessa comparacao podemos ver que o pretratamento destacou as principais

% bandas do espectro e suavizou a area de ruido. Sera que essa mudanca

% melhorou o nosso modelo? Vamos testar.

options=[];

options.Xpretreat = {'center'};

options.vene = 5;

options.vl = 20;

modelo=plsmodel2(Xcal2,ycal,options);

% Repare, caso tu nao tenha fechado a imagem do subplot, o octave/matlab

% fara por cima o grafico de RMSECV x VL. Mas isso nao nos atrapalha.

% Pelo grafico, podemos testar 6 e 8 variaveis latentes, antes, vamos

% tratar o Xtest.

[Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[15,2,1]});

options.vl = 6;

modelo6 = plsmodel2(Xcal2,ycal,Xtest2,ytest,options);

options.vl = 8;

modelo8 = plsmodel2(Xcal2,ycal,Xtest2,ytest,options);

%VL 6 8

%RMSEC 2.0981 1.7265

%RMSEP 1.3693 1.3686

%R2c 0.9532 0.9691

%R2p 0.9759 0.9727

% Nao e preciso usar o teste de acuracia para perceber que os modelos sao

% bem proximos, isso com base no RMSEP e R2p, apesar do RMSEC apontar uma

% boa diferenca o 6 tem um numero menor de variaveis latentes, entao,

% este e considerado o melhor modelo.

clear modelo8 %Excluindo para nao atrapalhar comparacoes futuras.

% Nao precisamos nos limitar a usar somente um pretratamento, podemos

% combinar pretratamentos.

[Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[15,2,1]});

[Xcal2,Xtest2]=pretrat(Xcal2,Xtest2,{'msc'});

% Lembre-se sempre de conservar a matriz X original.

options=[];

options.Xpretreat = {'center'};

options.vene = 5;

options.vl = 20;

modelo=plsmodel2(Xcal2,ycal,options);

% Vamos testar o VL = 9;

options.vl = 9;

modelo9 = plsmodel2(Xcal2,ycal,Xtest2,ytest,options);

%Pretrat deriv deriv/msc

%VL 6 9

%RMSEC 2.0981 0.8704

%RMSEP 1.3693 1.1094

%R2c 0.9532 0.9923

%R2p 0.9759 0.9816

% Ao olhar os parametros de avaliacao, podemos concluir que o pretratamento

% duplo e com 9 VL e o melhor modelo, todavia so podemos confirmar isso

% apos uma avaliacao minuciosa.

% Grafico Medido x Predito

close all %Fechando imagens ja criadas.

subplot(2,1,1)

plot(modelo6.Ycal(:,1),modelo6.Ycal(:,2),'bo','LineWidth',1); hold on;

plot(modelo6.Ytest(:,1),modelo6.Ytest(:,2),'r\*','LineWidth',1); hold on;

ylim([5 65]); xlim([5 65]);

plot(xlim, ylim, '--k');legend('Calibration','Prediction','Location','southeast');

title('Modelo 6');

set(gca,'FontSize',12);xlabel('Reference','fontsize',12);

ylabel('Predicted','fontsize',12);

subplot(2,1,2)

plot(modelo9.Ycal(:,1),modelo9.Ycal(:,2),'bo','LineWidth',1); hold on;

plot(modelo9.Ytest(:,1),modelo9.Ytest(:,2),'r\*','LineWidth',1); hold on;

ylim([5 65]); xlim([5 65]);

plot(xlim, ylim, '--k');legend('Calibration','Prediction','Location','southeast');

title('Modelo 9');

set(gca,'FontSize',12);xlabel('Reference','fontsize',12);

ylabel('Predicted','fontsize',12);

% Ao analisar os graficos, vemos que o modelo 9 conseguiu deixar as

% amostras mais proximas da linha de referencia em toda a faixa da

% concentracao.

% Teste de Acuracia.

tic

[pvalue,dist\_tt,meandiff] = accuracy\_test(ytest,modelo6.Ytest(:,2),modelo9.Ytest(:,2),'randbi',500000,0.05);

toc %{875 sec\_Pedro}

% A acuracia de ambos modelos sao estatisticamente semelhantes, entao, nao

% podemos afirmar que o modelo com duplo pretratamento e VL 9 e melhor

% estatisticamente, mas podemos escolhelo como melhor modelo do conjunto.

%% - Deteccao de Outlier

% Outlier, ou amostra anomala, ou amostra atipica, trata-se de uma amostra

% que esta anormalmente distante das demais amostras do conjunto, podendo

% se tratar tanto de uma amostra que nao se encaixa naquele conjunto

% amostral ( Uma amostra de oleo de canola no meio de amostras de azeite),

% um erro do procedimento da fonte analatica ( Um infravermelho com

% defeito) e um erro humano ( Erro na hora de digitar a concentracao da

% amostra.).

% Essas amostras anamalas tem a capacidade de prejudicar a modelagem

% aplicada, criando modelo com vies incorretos, prejudicando a otimizacao,

% e ate maquiando os parametros de avaliacao, desse modo, torna-se

% interessante remove-las da modelagem.

% Assim, iremos aprender a utilizar o grafico de Leverage e Residue para

% identificar outlier.

% Leverage e Residue

clearvars -except ycal ytest Xcal Xtest objetos Nmir modelo9 modelo6

[Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[15,2,1]});

[Xcal2,Xtest2]=pretrat(Xcal2,Xtest2,{'msc'});

modelo9 = lev\_res(modelo9,Xcal2,ycal,Xtest2,ytest);

close all

% A funcao lev\_res ira calcular tanto o leverage quanto o residuo das

% amostras, mas vamos analisar cada um isolamente e depois ambos em

% conjunto.

% O 'Leverage', nao tem um traducao correta, trata-se de uma medida que

% analisa a influencia de uma amostra na construcao do modelo de regressao.

% Um leverage baixo significa pouca influencia e um alto grande influencia.

% Caso uma amostra venha de um conjunto amostral diferente a tendencia e

% que ela tenha um grande leverage, destancando-se e prejudicando o modelo.

% Isso pode ocorrer, ou nao, por ser um outlier.

figure(1)

plot(1:1:size(modelo9.lev\_res.lev\_cal,1),modelo9.lev\_res.lev\_cal,'bo'); hold on;

plot(1:1:size(modelo9.lev\_res.lev\_test,1),modelo9.lev\_res.lev\_test,'r\*'); hold on;

hline(modelo9.lev\_res.lev\_limite,'k');

title('Leverage');

% No grafico podemos perceber que tem seis amostras que apresentam alta

% influencia no modelo, sendo a de maior influencia a amostra numero 71 da

% calibracao com 0,6111 de influencia, sendo que o limite calculado e

% 0,3253 (Linha divisoria). Todavia, so este teste nao e o suficiente para

% classificar as amostras com outlier e interessante analisar o residuo

% tambem.

figure(2)

for qi=1:1:size(modelo9.lev\_res.res\_cal,1)

plot(ycal(qi),modelo9.lev\_res.res\_cal2(qi),'bo'); hold on;

end

for qi=1:1:size(modelo9.lev\_res.res\_test,1);%qi=1

plot(ytest(qi),modelo9.lev\_res.res\_test2(qi),'r\*'); hold on;

end

title('Residue')

hline(0,'k:');

% Quando analisamos o grafico de residuo nao reparamos em nenhuma tedencia,

% ou anormalidade na distribuicao, assim podemos concluir que nao tem

% outliers indicados por esse teste.

% Por ultimo, iremos analisar o grafico combinado.

modelo9 = lev\_res(modelo9,Xcal2,ycal,Xtest2,ytest);

% Neste grafico combinamos a analise do Leverage com a analise do Residuo,

% caso uma amostra esteja no priemiro quadrante, quadro superior direito, e

% um forte indicio que ela seja um outlier. Neste modelo nao encontramos

% nenhuma amostra ali, tu pode mover a legenda caso precise. Mas sera que

% no modelo de VL 6 temos? Vamos verificar.

close all

[Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[15,2,1]});

modelo6 = lev\_res(modelo6,Xcal2,ycal,Xtest2,ytest);

% Vamos remover a legenda para melhor visualizar.

legend('off')

% Pelo grafico vemos que temos tres amostra de calibracao com forte indicio

% de outlier, entao vamos localiza-las e remove-las. Ao colocar o curso nas

% amostras, ou seleciona-la (matlab), podemos ver a coordenada.

% x = 0.274 / 0.278 / 0.338

% y = 2.657 / 2.693 / 2.933

A = find( 0.273 < modelo6.lev\_res.lev\_cal)

% Com esta funcao encontraremos a amostra com leverage em calibracao maior

% que 0.273. Amostra numero 25,31 e 56. Como se trata de amostras no conjunto

% calibracao, teremos que refazer todo o processo de otimizacao e previsao.

% Removendo as amostras da calibracao.

Xcal\_2 = Xcal; ycal\_2 = ycal;

Xcal\_2(A,:) = []; ycal\_2(A,:) = [];

% Recomendo sempre guarda a versao original do conjunto amostral, ao

% remover uma amostra, simplesmente crie uma copia e adicionar "\_2" no

% nome, para nao ter confusao.

% Pretratamento

[Xcal2\_2,Xtest2]=pretrat(Xcal\_2,Xtest,{'deriv',[15,2,1]});

% Preprando o PLS

options=[];

options.Xpretreat = {'center'};

options.vene = 20;

options.vl = 20;

modelo6\_2=plsmodel2(Xcal2\_2,ycal\_2,options);

% Escolhendo VL = 7, como melhor.

options.vl = 7;

modelo6\_2=plsmodel2(Xcal2\_2,ycal\_2,Xtest2,ytest,options);

%Outlier Com Sem

%VL 6 7

%RMSEC 2.0981 1.8091

%RMSEP 1.3693 1.2058

%R2c 0.9532 0.9621

%R2p 0.9759 0.9702

% Pelos parametros de avalicao os modelos nao tiveram uma grande diferenca.

% Vamos utilizar o teste da acuracia e verificar se tem diferenca

% significativa.

tic

[pvalue,dist\_tt,meandiff] = accuracy\_test(ytest,modelo6.Ytest(:,2),modelo6\_2.Ytest(:,2),'randbi',500000,0.05);

toc {868 seg\_Pedro}

% A diferenca nao foi significativa, vamos para o teste do lev\_res.

modelo6\_2 = lev\_res(modelo6\_2,Xcal2\_2,ycal\_2,Xtest2,ytest);

legend('off')

% Dessa vez uma amostra de teste foi indetificada com indicio de ser

% outlier, lembre-se um outlier para um modelo nao quer dizer outlier para

% outro. Vamos remover essa amostra teste e fazer, somente, a parte do

% teste.

A = find( 4 < modelo6\_2.lev\_res.res\_test);

% Encontando a amostra teste. Amostra 9.

Xtest\_2 = Xtest; ytest\_2 = ytest;

Xtest\_2(9,:) = []; ytest\_2(9,:) = [];

% Removendo o outlier

[Xcal2\_2,Xtest2\_2]=pretrat(Xcal\_2,Xtest\_2,{'deriv',[15,2,1]});

options.vl = 7;

modelo6\_2=plsmodel2(Xcal2\_2,ycal\_2,Xtest2\_2,ytest\_2,options);

% Refazendo o modelo.

%Outlier Com Sem

%VL 6 7

%RMSEC 2.0981 1.8091

%RMSEP 1.3693 1.0053

%R2c 0.9532 0.9621

%R2p 0.9759 0.9808

% Agora temos uma mudanca significativa nos parametros de avalicao do

% teste.

tic

[pvalue,dist\_tt,meandiff] = accuracy\_test(ytest,modelo6.Ytest(:,2),modelo6\_2.Ytest(:,2),'randbi',500000,0.05);

toc

% Note que dara um erro, o teste de acuracia so funciona quando o conjunto

% ytest tem a mesma quantidade de amostras, como removemos uma amostra, nao

% podemos utilizar este teste. Entao temos que utilizar um teste nao

% parametrico, todavia, como trata-se de um assunto complexo deixaremos

% isso para outro tutorial.

% Refazendo o teste de outlier.

modelo6\_2 = lev\_res(modelo6\_2,Xcal2\_2,ycal\_2,Xtest2\_2,ytest\_2);

legend('off')

% Agora nao temos nenhum amostra anomola.

%% Selecao de Variaveis

% O aperfeiçoamento por Selecao de Variaveis pode ser realizado por

% diversas tecnicas, o seu foco e conseguir selecionar as variaveis da

% fonte analitica que tem as informacoes mais importantes para o modelo e

% como consequencia, remover variaveis com pouca, ou nenhuma, informacao,

% como ruido, e aperfeicoar o modelo.

% Infelizmente, como se trata de um assunto complexo e cheio de detalhes,

% nao iremos aplicar/ensinar neste tutorial, mas o faremos num proximo.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

%%

%%%%%%%%%%%%%% 03 - Exemplo real.

% Aqui iremos simular um pratica da quimiometria envolvendo o PLS. Como se

% fossemos um quimico recebendo o dado e precisando trata-lo com

% quimiometria.

clear all;

clc;

close all;

% Agora iremos puxar os pacotes necessarios. (OC)

pkg load statistics

pkg load io

cd('C:\Users\Pedro\OneDrive - aluno.ufes.br\Quimiometria\IFES Ciencia\PLS\_Model');

%cd('...\IFES Ciencia\PLS\_Model');

% Como extrair dados de planilha excel.

[y,~,~]=xlsread('Oleos\_Adulterados.xlsx','Plan1','B2:B230');

%[A,B,C]=xlsread('XXX','YYY','ZZZ');

% A funcao xlsread e utilizada para extrair informacao de planilhas como

% xls, xlsx e csv.

% INPUT:

% XXX = O nome da planilha desejada, incluindo o formato.

% YYY = A aba que os dados estao.

% ZZZ = A faixa que os dados se encontram.

% OUTPUT:

% A = Caso o dado seja numero.

% B = Caso o dado seja caracteres.

% C = Numerico e letras.

[num,~,~]=xlsread('Oleos\_Adulterados.xlsx','Plan1','C1:DW1');

[X,~,~]=xlsread('Oleos\_Adulterados.xlsx','Plan1','C2:DW230');

[~,Sample,~]=xlsread('Oleos\_Adulterados.xlsx','Plan1','A2:A230');

% Estes dados sao dados de Azeite adulterados com oleo, provindos do artigo

% [10.1016/j.focha.2022.100074], coletadas com autorizacao e modificados

% para nao ficar identico.

% Algumas pessoas maliciosas tendem a falsificar azeite com oleos baratos

% com o objetivo de ter um alto lucro enquanto prejudicam o consumidor.

% Assim, artigos como esse, que aplicam uma tecnica portatil de

% infravermelho, no caso o infravermelho proximo, sao de suma importancia.

% A primeira coisa que se deve fazer ao receber um dado desse e analisar o

% espectro.

plot(num,X);

% Note que temos uma amostra anomala no conjunto, nitidamente ela nao tem o

% mesmo perfil espectroscopico das demais amostras, assim, e interessante

% ja remover restas amostra, afinal, tratase de um outlier.

AAA= find(X(:,1) > 1);

X(AAA,:) = []; y(AAA,:) = [];

%% Separacao Cal Test

% Agora vamos aprender a separar as amostras em conjunto calibracao e

% teste, que ate entao, as amostras ja vinham separadas corretamente.

[objetos,Xcal,Xtest,ycal,ytest]=caltest(X,y,70,'k',0,{'center'});

% A funcao caltest foi criada com o objetivo de separar conjuntos amostrais

% de regressao em conjunto calibracao e conjunto teste, existe diversos

% detalhes nele como podemos ver abaixo.:

%[A,B,C,D,E]=caltest(XXX,YYY,ZZZ,WWW,VVV,UUU);;

% A funcao xlsread e utilizada para extrair informacao de planilhas como

% xls, xlsx e csv.

% INPUT:

% XXX = Fonte analitica das amostras.

% YYY = Vetor informativo dos dados de regressao.

% ZZZ = Percentagem que deve esta no conjunto calibracao. (Recomendo 70)

% WWW = Algoritmo desejado. (Recomendo 'k' kenston)

% VVV = Checar repeticao. (0 = Sem checagem 1 = Com checagem)

% UUU = Metodo de pretratamento utilizado antes da separacao. (Recomendo

% 'none')

% OUTPUT:

% A = Objetos, conjunto estrutural com os dados da separacao. (Recomendo

% sempre salvar)

% B = Fonte Analitica conjunto calibracao.

% C = Fonte Analitica conjunto teste.

% D = Vetor informativo conjunto calibracao.

% E = Vetor informativo conjunto teste.

% Apos a separacao e interessante analisar como o conjunto foi separado,

% utilizando os seguintes comandos.

close all

plot(1:1:size(ycal,1),ycal,'bo'); hold on;

plot(1:1:size(ytest,1),ytest,'r\*'); hold on;

% O mais imporante aqui e perceber se no cojunto calibracao, em bolinhas

% azuis, estao as amostras de menor e maior valor, para o conjunto teste

% esta dentro desta faixa.

% Como percebido, no conjunto teste tem amostras com o mesmo valor minimo

% do conjunto calibracao, todavia, isso nao e problematico. Todavia, este

% conjunto tratase de uma triplicata, como o Sample indica.

Sample([1 2 3],:) % Para ver o nome das primeiras amostras.

% Vemos que as 3 primeiras amostras sï¿½o triplicatas nao reais, entao, por

% REGRA, elas tem que estar no mesmo conjunto, seja calibracao, seja teste.

% O que nao ocorre quando analisamos o objetos;

objetos.cal([1 2 3],:) % Para visualizar as amostras em calibracao.

% Que sao as amostras 1 3 e 5.

objetos.test([1 2 3],:) % Para visualizar as amostras em teste.

% Que sao as amostras 2 4 e 7.

% Como (1,2 e 3) deveriam estar no mesmo conjunto, essa separacao nao e

% valida. O mesmo vale para o conjunto (4,5 e 6).

% Desse modo, nao poderemos fazer uso dessa separacao do caltes, mas, caso

% fizesemos a media das amostras, poderiamos. Aqui, vamos fazer a mesma

% separacao que o artigo usou, com base na concentracao de adulterante.

% Para isso iremos combinar for e if. Para melhor compreensao do for, peco

% que leiam o "00 - Conhecendo a funcao for", afinal, este tutorial ja esta

% grande.

objetos.cal = [];

objetos.test = [];

for qi=1:1:size(X,1);

% Se y tem valor 3 7 10 e 30;

if y(qi) == 3 || y(qi) == 7 || y(qi) == 10 ||y(qi) == 30;

objetos.test = [objetos.test;qi];

else

objetos.cal = [objetos.cal;qi];

end

end

Xcal = X(objetos.cal,:); Xtest = X(objetos.test,:);

ycal = y(objetos.cal,:); ytest = y(objetos.test,:);

close all

plot(1:1:size(ycal,1),ycal,'bo'); hold on;

plot(1:1:size(ytest,1),ytest,'r\*'); hold on;

% Note que esta separacao segue todos pre-requisitos citados. Agora vamos

% para a modelagem.

%% Modelando

close all

options=[];

options.Xpretreat = {'center'};

options.vene = 5;

options.vl = 20;

modelo=plsmodel2(Xcal,ycal,options)

options.vl = 9;

modelo9 = plsmodel2(Xcal,ycal,Xtest,ytest,options);

options.vl = 15;

modelo15 = plsmodel2(Xcal,ycal,Xtest,ytest,options);

%VL 9 15

%RMSEC 4.1114 3.4256

%RMSEP 5.0965 4.6783

%R2c 0.9377 0.9584

%R2p 0.7597 0.7937

% O VL melhor foi o 15 como podemos analisar com base nos parametros de

% avaliacao, todavia, como estamos falando de deteccao de adulterantes e

% importante analisarmos outros parametros de avaliacao. Limite de Deteccao

% e Limite de Quantificacao, e interessante conserguimos detectar o menor

% valor possivel de adulterante em cada amostra. Assim, utilizamos uma nova

% funcao.

modelo9=eparamenter(modelo9,Xcal,Xtest,ycal,ytest);

modelo15=eparamenter(modelo15,Xcal,Xtest,ycal,ytest);

% Esta funcao foi desenvolvida para ser utilizada em conjunto com o

% plsmodel, assim, precisa do input "modelo" e dos demais inputs para

% funcionar, em resposta ela adiciona uma nova estrutura dentro do

% "modelo". Com diversos novos parametros de avaliacao.

%VL 9 15

%RMSEC 4.1114 3.4256

%RMSEP 5.5650 5.4313

%R2c 0.9377 0.9584

%R2p 0.7597 0.7937

%LD 5.5375 7.1088

%LQ 18.4584 23.6959

% Note que ao analisar o LD o modelo com 9 variveis latentes demonstrou um

% resultado melhor, entao neste caso o melhor modelo seria o VL 9.

% Todavia, tem dois testes que ainda precisamos fazer para aprovar este

% modelo.

%Teste de bias

modelo9.bias.c=bias\_teste(ycal,modelo9.Ycal(:,2),0.05);

modelo9.bias.p=bias\_teste(ytest,modelo9.Ytest(:,2),0.05);

% O teste de Bias visa analisar erros sistemï¿½ticos no modelo, caso tenha

% erro, o modelo deve ser descartado e um com menos VL deve ser feito

% (Cuidado). Aqui no caso o nosso modelo foi aprovado, todavia, esta longe

% de ser o melhor modelo.

% Analisando o medido de previsto.

plot(modelo9.Ycal(:,1),modelo9.Ycal(:,2),'bo','LineWidth',1); hold on;

plot(modelo9.Ytest(:,1),modelo9.Ytest(:,2),'r\*','LineWidth',1); hold on;

ylim([0 65]); xlim([0 65]);

plot(xlim, ylim, '--k');legend('Calibration','Prediction','Location','southeast');

title('Modelo 9');

set(gca,'FontSize',12);xlabel('Reference','fontsize',12);

ylabel('Predicted','fontsize',12);

% Note que 5 amostras, duas na concentracao 7, uma na concentracao 9, uma

% na 10 e uma 30, se separaram drasticamente da linha de referencia, isso e

% um indicio de outlier, e nenhuma concentracao conseguiu ficar bem proxima

% da linha de referencia em conjunto.

% Vamos testar Outlier...

modelo9 = lev\_res(modelo9,Xcal,ycal,Xtest,ytest);

close all

figure(1)

plot(1:1:size(modelo9.lev\_res.lev\_cal,1),modelo9.lev\_res.lev\_cal,'bo'); hold on;

plot(1:1:size(modelo9.lev\_res.lev\_test,1),modelo9.lev\_res.lev\_test,'r\*'); hold on;

hline(modelo9.lev\_res.lev\_limite,'k');

title('Leverage');

% Temos 10 amostras acima da linha limite, uma drasticamente afastada, um

% forte indicio de outlier.

figure(2)

for qi=1:1:size(modelo9.lev\_res.res\_cal,1)

plot(ycal(qi),modelo9.lev\_res.res\_cal2(qi),'bo'); hold on;

end

for qi=1:1:size(modelo9.lev\_res.res\_test,1);%qi=1

plot(ytest(qi),modelo9.lev\_res.res\_test2(qi),'r\*'); hold on;

end

title('Residue')

hline(0,'k:');

% Ignorando 4 amostras abaixo de -10, o residuo parece aumentar conforme a

% concentracao aumenta, isso pode ser ou um erro de tendencia ou devido o

% aumento da percentagem de oleo.

figure(3);

modelo9 = lev\_res(modelo9,Xcal,ycal,Xtest,ytest);

% Como suspeito, temos 3 amostras que podem ser consideradas outlier, ambas

% no conjunto teste, vamos remove-las e ver como fica nosso modelo.

A = find( 0.2750 < modelo9.lev\_res.lev\_test); Procurando amostras teste

% com leverage acima de 0.2824. Note que peguamos 4 amostras, sendo so tres

% sao outlier, isso ocorre porque tem um nao outlier que se encontra na

% faixa que escolhemos. Entao, temos que utilizar 2 condicoes para

% encontralo com precisao.

A = find( 0.2750 < modelo9.lev\_res.lev\_test & 11.13< modelo9.lev\_res.res\_test)

% Agora sim, os 3 outliers. [5,47,54]

% Como se trata de suspeita de outlier em conjunto test, refazemos sï¿½ a

% parte do teste.

Xtest\_2 = Xtest; ytest\_2 = ytest;

Xtest\_2(A,:) = []; ytest\_2(A,:) = [];

options.vl = 9;

modelo9\_2 = plsmodel2(Xcal,ycal,Xtest\_2,ytest\_2,options);

modelo9\_2 = eparamenter(modelo9\_2,Xcal,Xtest,ycal,ytest);

%VL 9 9\_2

%RMSEC 4.1114 4.1114

%RMSEP 5.5650 4.2089

%R2c 0.9377 0.9377

%R2p 0.7597 0.8431

%LD 5.5375 5.5375

%LQ 18.4584 18.4584

% Com estes parametros de avaliacao, podemos dizer que houve uma melhora no

% modelo.

%Teste de bias

modelo9\_2.bias.c=bias\_teste(ycal,modelo9\_2.Ycal(:,2),0.05);

modelo9\_2.bias.p=bias\_teste(ytest\_2,modelo9\_2.Ytest(:,2),0.05);

% Analisando o medido de previsto.

plot(modelo9\_2.Ycal(:,1),modelo9\_2.Ycal(:,2),'bo','LineWidth',1); hold on;

plot(modelo9\_2.Ytest(:,1),modelo9\_2.Ytest(:,2),'r\*','LineWidth',1); hold on;

ylim([0 65]); xlim([0 65]);

plot(xlim, ylim, '--k');legend('Calibration','Prediction','Location','southeast');

title('Modelo 9');

set(gca,'FontSize',12);xlabel('Reference','fontsize',12);

ylabel('Predicted','fontsize',12);

% Analise de leverage e residue, nao precisa ser refeita porque so mexemos

% com o teste, caso ocorrese uma exclusao de amostra de calibracao, todo o

% processo precisaria ser refeito.

%%

%%%%%%%%%%%%%% 04 - Exercicio para praticar.

% Agora, exericio para praticar, no conjunto do plsmodel2 tem um workspace

% chamaro "Nitrogenio Total.mat", com as seguintes informacoes:

% Indet : Identificacao das amostras.

% num : Numero de onda do espectro de infravermelho.

% X : Espectro de Infravermelh Medio das amostras. [Fonte Analitica]

% y : Vetor de concentracao de Nitrogenio total.

% Alem disso, os modelos utilizados como exemplo aqui neste tutorial nao

% sao os melhores obtidos pelo nosso laboratorio, entao se sinta desafiado

% a tentar encontra-los. O Gabarito se encontra no final desta rotina.

%%

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

%%%%%%%%%%%%%% 00 - Edicao de figuras

% Neste tutorial sera ensinado como editar configuracoes das imagens do

% Octave.

% {AINDA SERA FATO}

%%

%%%%%%%%%%%%%% 00 - Edicao de Figuras da Monografia.

%%

%%

%%%%%%%%%%%%%% 00 - Conhecendo a funcao for

%%

%%%%%%%%%%%%%% 00 - Gabarito

% 01-02

%VL 12

%RMSEC 0.7243

%RMSEP 0.8121

%R2c 0.9949

%R2p 0.9866

%LD 0.3422

%LQ 1.1407

%Pret DerivNorm

% 03

%VL XX

%RMSEC XXX

%RMSEP XXX

%R2c XXX

%R2p XXX

%LD XXX

%LQ XXX

%Pret XXX

% 04

%VL XX

%RMSEC XXX

%RMSEP XXX

%R2c XXX

%R2p XXX

%LD XXX

%LQ XXX

%Pret XXX